Approaches, Challenges, and Applications for Deep Visual Odometry: Toward to Complicated and Emerging Areas

Ke Wang, *Member, IEEE,* Sai Ma, Junlan Chen, Jianbo Lu, *Fellow, IEEE*

***Abstract*—Visual odometry (VO) is a prevalent way to deal with the relative localization problem, which is becoming increasingly mature and accurate, but it tends to be fragile under challenging environments. Comparing with classical geometry-based methods, deep learning-based methods can automatically learn effective and robust representations, such as depth, optical flow, feature, ego-motion, etc., from data without explicit computation. Nevertheless, there still lacks a thorough review of the recent advances of deep learning-based VO (Deep VO). Therefore, this paper aims to gain a deep insight on how deep learning can profit and optimize the VO systems. We first screen out a number of qualifications including accuracy, efficiency, scalability, dynamicity, practicability, and extensibility, and employ them as the criteria. Then, using the offered criteria as the uniform measurements, we detailedly evaluate and discuss how deep learning improves the performance of VO from the aspects of depth estimation, feature extraction and matching, pose estimation. We also summarize the complicated and emerging areas of Deep VO, such as mobile robots, medical robots, augmented and virtual reality, etc. Through the literature decomposition, analysis, and comparison, we finally put forward a number of open issues and raise some future research directions in this field.**

***Index Terms*—Visual Odometry, Deep Learning, Pose Estimation, Motion Estimation, SLAM.**

I. Introduction

VISUAL odometry is the problem of estimating the camera pose from consecutive images and is a fundamental capability required in many computer vision and robotics applications, such as Cognitive Robots, Autonomous and Evolutionary Robots, Medical Robots, Augmented / Mixed / Virtual Reality and other complicated and emerging applications based on localization information, such as indoor and outdoor navigation, scene understanding, space exploration [1]–[3].

Over the past decades, we have seen impressive progress on the classical geometry-based visual odometry. Superior performance on accurate and effective are demonstrated in both feature-based methods [4]–[6] and direct methods [7]–[9], which pushes the classical geometry-based visual odometry towards real-world applications [10]–[14]. However, the robustness of these methods cannot meet the requirements of high-reliability robots under challenging environments, such as illumination changing and dynamic environments, etc.

TABLE I
THE CLASSICAL ALGORITHM VERSUS THE DEEP LEARNING ALGORITHM

	Classical	**Deep learning**
Representation	Explicit: visual geometric model	Implicit: network structure and parameter
Accuracy	High: static environments	High: realistic environments
Robustness	Bad: need excellent models	Good: trained regime
Efficiency	High: only few parameter	Low: Huge number of parameters
Generalization	Widely applicable	Only in trained region
Data	Only needed for system initialization	Training needs large-scale data
Application scenarios	Static, texture sufficiency, no illumination changing	Texture-less, changing environments

Recently, the development of mobile devices especially smart-phone make visual odometry more accessible for common users, which derivatives many emerging applications and also brings many challenges. Unfortunately, relying solely on classical geometry-based methods cannot process these challenges, as these methods are highly dependent on the low-level hand-designed features that do not represent well the complex environment in the real world. Besides, the geometry-based method assume that the world is static. In practice, the scene geometry and appearance of the real-world changes significantly over time. Deep learning shows a powerful capability in many computer vision tasks. This prompted the researcher to think about the possibility of applying deep learning to visual odometry. Table I summarizes the main properties of classical algorithms and deep learning-based algorithms.

Comparing with the classical geometry-based method, deep learning can automatically learn more effective and robust feature representations without manual feature design when trained on sufficiently large-scale datasets. It seems that the interest point extracted by learning model is more robust to changing environments. Although these learning-based models have many deficiencies and inaccuracies, they can continue to learn and adapt to new environments, just as humans rely on intuition rather than accurate models to operate. Besides, by integrating the knowledge of geometry and visual psychophysics, it becomes possible for the system to acquire

K. Wang and S. Ma are with the School of Automobile Engineering, Chongqing University, China, 400044, also with the Key Lab of Mechanical Transmission, Chongqing University, China, 400044 (e-mail: kw@cqu.edu.cn, masai@cqu.edu.cn).

J. Chen is with shool of Economics and Management, Chongqing Normal University, Chongqing 401331, China (e-mail: nwpujunlan@163.com).

J. Lu is with Research and Advanced Engineering, Ford Motor Company, Dearborn, MI 48121 USA (e-mail: jlu10@ford.com).

Manuscript received April 19, 2005; revised August 26, 2015. (Corresponding author: Ke wang and Junlan Chen)

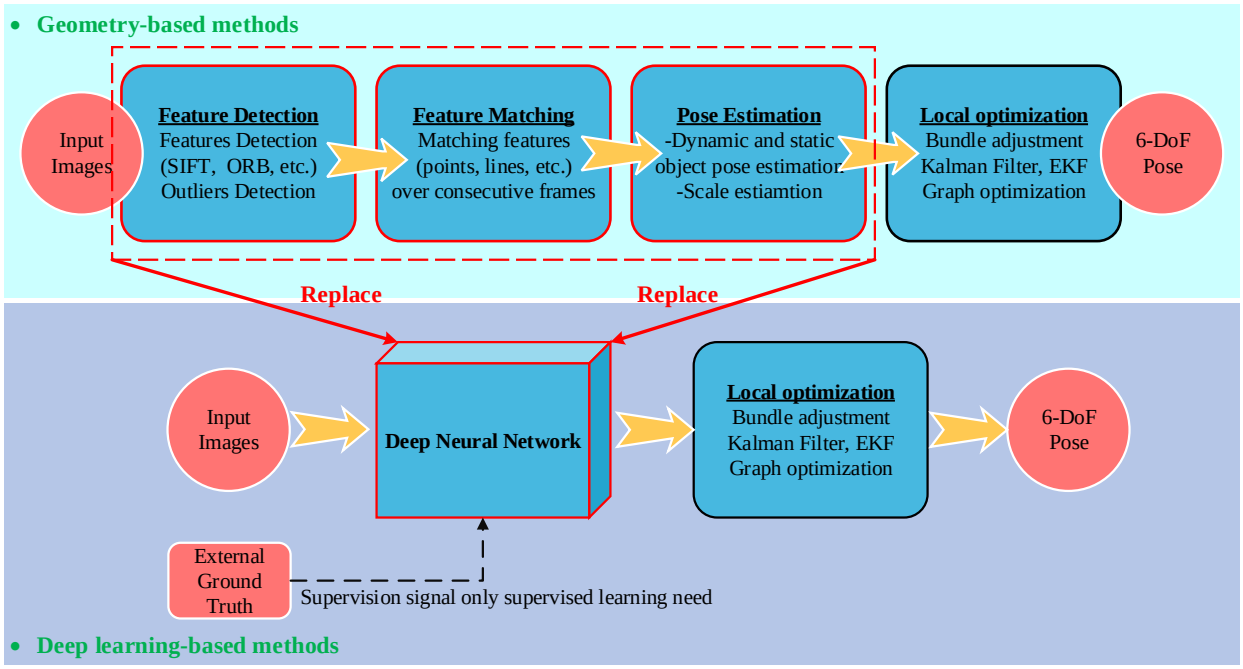

Fig. 1. Geometry-based and Deep learning-based visual odometry paradigms. The geometry-based visual odometry computes the camera pose from the image by extracting and matching feature points. The deep learning-based visual odometry can estimate the camera pose directly from the data. For supervised visual odometry, it requires external ground truth as a supervision signal, which is usually expensive. In contrast, unsupervised visual odometry does not require this. Besides, the local optimization module is optional for deep learning-based visual odometry.

a high-level understanding of the environment and achieve active(task-driven) perception. Many advanced researches have shown that learning-based methods can achieve more robust performance under certain specific conditions, and have demonstrated the huge potential of integrated deep learning to solve the problems faced by geometry-based methods [23]–[26]. Therefore, there is an unavoidable trend to use learning-based methods in visual odometry to improve the performance. Fig.1 shows the comparison of the geometry-based and deep learning-based visual odometry pipeline.

As shown in table II, there are some review papers about visual odometry or SLAM. Most recent review papers focus on geometry-based methods, less on deep learning-based approaches [15], [17], [18]. Some review papers focus on specific aspects including multi-robot SLAM [19], semantic information [22], and dynamic environment processing [21]. However, these review papers not mainly focus on visual odometry. Aqel et al. [16] was the only review that discussing visual odometry separately. But this review only focus on the geometry-based methods. Finally, Li et al. [20] provided a review that focuses the ongoing evolution of visual SLAM from model-based methods to learning-based methods. Therefore, they only reviewed the early work of learning-based methods. However, deep learning-based methods have developed rapidly and have achieved significant progress in recent years.

In summary, we can find that most existing review papers focus on SLAM in the perspective of geometry. A review specific to visual odometry based on deep learning is still lacking. Therefore, in this paper, we provide a review on deep learning-based visual odometry including techniques, applications, open questions and opportunities. **In particular, our main contributions are as follows:**

- We propose a list of criteria as a uniform benchmark to evaluate the performance of the algorithm.
- We review the literature of visual odometry based on deep learning in three aspects by using the offered criteria as the uniform measurements. Then, we analyze and discuss the advantages and challenges of these methods in detail. Finally, how deep learning can improve the performance of visual odometry is also commented.
- We sum up the application of visual odometry and point out that Deep VO will play a big role in complicated and emerging areas.
- Based on the completed review and in-depth analysis, we put forward a number of open problems and raise some opportunities.

The rest of this paper is organized as follows: In section II, a list of criteria is proposed, which serve as a uniform benchmark to evaluate the performance of the algorithms. In section III, we review the literature in three aspects based on the proposed criteria. In section IV, we sum up the application of visual odometry. Challenges and opportunities are given in section V, followed by a conclusion in section VI.

II. Criteria of Deep Learning for Visual Odometry

In this section, we propose a number of criteria for deep learning-based visual odometry to evaluate the literature in a uniform manner in the next section. We also discussed how to select the algorithms according to the evaluation results. The notations used to evaluate these algorithms according to proposed criteria are list in Table III.
The proposed criteria are described below:

TABLE II
SUMMARIZATION OF A NUMBER OF RELATED SURVEYS SINCE 2015

No	Survey Title	Ref	Year	Content
1	Visual Simultaneous Localization And Mapping: A Survey	[15]	2015	Odometry Geometry-based Methods
2	Review of Visual Odometry: Types, Approaches, Challenges, and Applications	[16]	2016	Odometry Geometry-based methods
3	Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age.	[17]	2016	Metric and Semantic SLAM Open problems and future direction
4	A Critique of Current Developments in Simultaneous Localization And Mapping	[18]	2016	Feature-based SLAM Pose graph SLAM using estimation and optimization techniques
5	Multiple-Robot Simultaneous Localization and Mapping: A Review.	[19]	2016	SLAM Miltiple Robot
6	Ongoing Evolution of Visual SLAM from Geometry to Deep Learning: Challenges and Opportunities.	[20]	2018	SLAM Geometry-based approach Deep learning-based approache
7	Visual SLAM and Structure from Motion in Dynamic Environments: A Survey.	[21]	2018	SLAM Structure from Motion In dynamic environments
8	Simultaneous Localization and Mapping in the Epoch of Semantics: A Survey.	[22]	2019	SLAM Semantic concept

TABLE III
THE NOTATION USED TO EVALUATE LEARNING ALGORITHMS ACCORDING TO PROPOSED CRITERIA

criteria	High or Yes	Medial	Low or No
Accuracy	Low RMSE translation and rotation drift with enough experiments	High RMSE translation or rotation drift without enough experiments	High RMSE translation and rotation drift without enough experiments
Efficiency	run in real-time; use resources economically and effectively		not run in real-time; not using resources economically and effectively
Scalability	work well in large-scale or long-term environments		poor work in large-scale or long-term environments
Practicability	make a trade-off between efficiency and accuracy; tested in practice		no trade-off between efficiency and accuracy only tested in datasets
Dynamicity	work well in dynamic environments		poor work in dynamic environments
Extensibility	easily improved; widely used;		difficult improved; rarely used;

(1) Accuracy: refers to how accurate localization is. It is the most important criteria for evaluating learning-based visual odometry models or algorithms. We divide the literature into two categories: the algorithm with specific results and the algorithm without specific results. For the former, we mainly use RMSE (Root Mean Square Error, defined in [27]) or ATE (Average Trajectory Error, define in [4]) to measure the bias of ground truth and estimation, which directly describes algorithm accuracy. Besides, for the literature without specific results, the different problem needs specific analysis.

(2) Efficiency: refers to how the algorithm use resources economically and effectively. For different application scenarios, system resources are limited in many aspects, such as computation, system memory. Using as less as possible resources to process as more as possible data in as shorter as possible time is a universal goal for learning-based visual odometry algorithms. We use computation time to reflect the algorithm efficiency to demonstrate algorithm advance through comparison with other algorithms.

(3) Scalability: refers to the ability of the algorithm to treat large-scale or long-term environments. For autonomous vehicles or other robots that works in large-scale outdoor environments, the algorithm should be able to work in continuous time with finite system memory. Further, the algorithm should find sufficiently discriminative features to against the changes (season, weather, trees leaves, etc.) in long-term environments.

(4) Dynamicity: refers to how the algorithm is able to deal with dynamic environments. Visual odometry techniques usually based on the assumption that the observed environment is static. The dynamic object can cause the wrong pose estimation. Therefore, alleviating the impact of dynamic objects is an effective but challenging approach to improve performance.

(5) Practicability: refers to whether the algorithm can be used in practical applications. To deploy an algorithm, it usually considers many aspects such as efficiency, accuracy, and extensibility. It includes the trade-off between efficiency and accuracy. It also includes easy further improved and widely used.

(6) Extensibility: refers to how algorithms or models can be easily further improved and widely used in various situations. First, the algorithm can automatically update the latest or temporary changes, and easy to maintain. Finally, the algorithm or

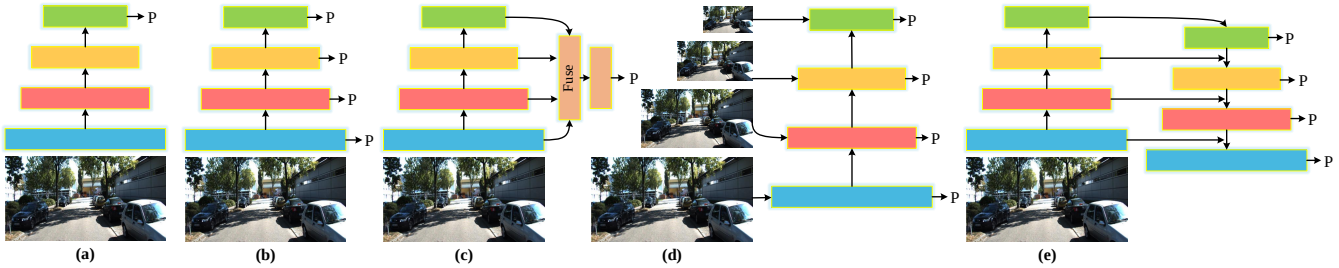

Fig. 2. An illustration and comparison of the paradigm for multi-scale feature learning used in depth estimation. (a) No multi-scale feature learning method is employed. (b) Depth is predicted on each feature map. (c) Depth is predicted on a single feature map, which is generated from multiple feature maps. (d) Depth is predicted from different scale images. (e) Before predicting depth on each feature map, the intermediate features from different scales are combined.

model trained in a specific environment that can be generalized well to 'unseen' environments without fine-tuning.

Based on the proposed criteria, we hope that these evaluation results can help us choose a suitable algorithm for a specific problem. Due to many robotic systems have strong computational and temporal restrictions, high accuracy and high efficiency usually not available at the same time. High accuracy usually means low efficiency. This is because the high accuracy network structure is complex and contains more parameters, which leads to longer inference time. Therefore, we should make a trade-off between accuracy and efficiency when choosing an algorithm for a specific application. If we have enough computing power, we can choose algorithms with high accuracy, and when the computing power is limited, we can choose algorithms with high efficiency or practicability. Then, we choose the algorithm according to the specific operating environment. If our system running in a dynamic, large-scale or long-term environment, we should choose the algorithm with scalability and dynamicity. Finally, in order to facilitate maintenance and upgrades, we usually choose the algorithm with extensibility.

III. Deep Learning for Visual Odometry

A. Deep learning for depth estimation

Depth estimation is the beginning of the combination of deep learning and visual odometry. These methods can be divided into two categories: supervised and unsupervised learning methods. Fig.3 shows the different neural network architectures used for depth estimation.

1) Supervised learning methods: For supervised learning, the use of supervised-based methods was first proposed in [28]. They proposed a multi-scale neural network for depth estimation from a single image. The network architecture consists of two components: one for coarse global predictions and one for refined local predictions. Since then, a lot of work has utilized the multi-scale features of the network and depth structure information to improve the performance of depth estimation [29]–[34]. Fig.2 is an illustration and comparison of the paradigm for multi-scale feature learning in depth estimation.

For depth estimation problem definition, previous work usually makes depth estimation as a regression problem. Therefore, Liu et al. [35] proposed a continuous CRF and CNN based framework similar to Fig.3(a), which transforms the depth estimation into learning problems. Besides, depth estimation can also be treated as a classification problem [36]. The intuition behind this is that it is easier to estimate the depth range than to estimate a specific value.

As far as we know, using CNN alone cannot model the long-range context well. Therefore, Grigorev et al. [37] proposed a hybrid network by combining convolutional layer and ReNet layer. The ReNet layer consists of Long Short-Term Memory units (LSTMs), so the ReNet layer can obtain a global context feature representation. Its network structure is similar to Fig.3(c). Moreover, an LSTM-based architecture was also proposed to explore whether RCNN can learn accurate spatial-temporally depth estimation without inter-frame geometric consistency or pose supervision [38].

To obtain more accurate and robust results, the researchers turned their attention to more challenging and specific problems. To alleviate the depth ambiguity problem, invariant surface normal was used to assist depth estimation [39]. And the ambiguity between focal length and monocular depth was demonstrated in [40]. Ma et al. [41] used both additional sparse depth samples acquired with a low-resolution depth sensor or computed via SLAM algorithms and input images to estimate the full-resolution depth. Additionally, a novel progressive hard-mining network was proposed in [42] to solve the problem of large semantic gap and accumulated error over scales. DFineNet was proposed in [43] to refine the depth estimation under sparse and noisy conditions.

2) Unsupervised learning methods : Although the supervised learning-based method has achieved promising results in depth estimation tasks, the major problem faced by supervised learning is that it requires explicit ground-truth information that is difficult to collect in some scenarios. Moreover, the collected dataset unavoidable exist errors because of the sensor accuracy limitation. The method using unsupervised learning was first introduced in [44], which estimates the depth information from single view images. This framework is analogous to an autoencoder and uses stereo images, in which the convolutional encoder takes the left image to predict the inverse depth map, and the decoder reconstructs the left image by synthesizing the right image with the predicted depth map. Similarly, Godard et al. [45] introduced a method that can simultaneously infer disparities that warp the left images to match the right one or warp the right image to match the left one, thereby obtaining better depth estimation by

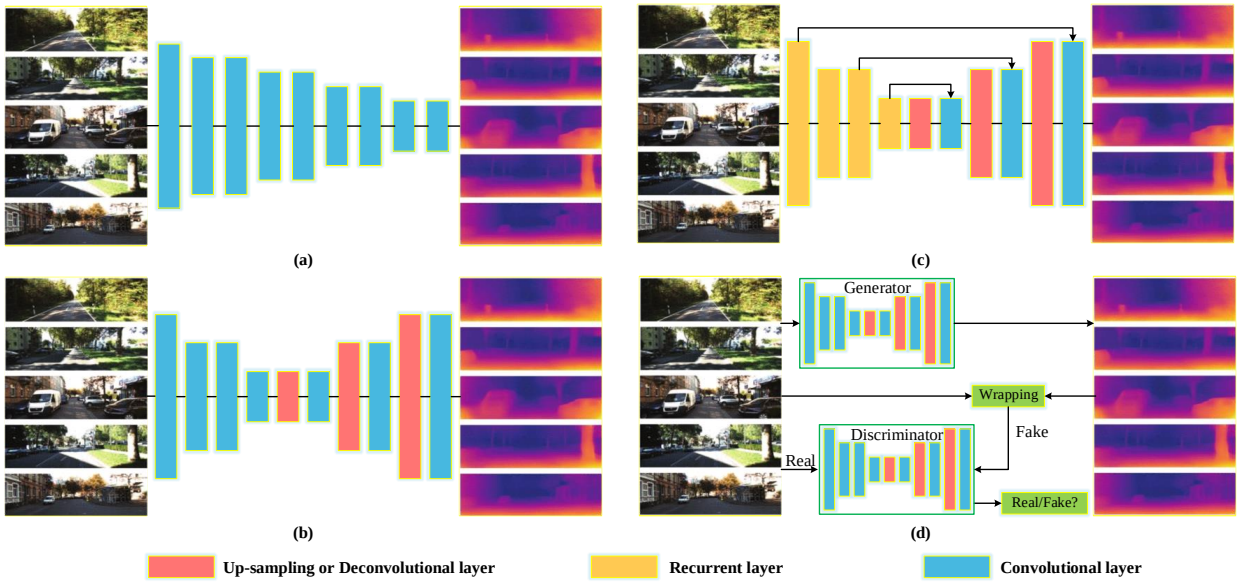

Fig. 3. An illustration of different neural network architectures for depth estimation. (a) CNN-based framework, which cannot obtain full resolution depth map. (b) CNN + Up-sampling or De-CNN based framework, which can obtain full resolution depth map. (c) RCNN based framework, which can obtain forward-backward consistent depth map. (d) GAN based framework.

enforcing left-right depth consistency. Zhou et al. [46] first synthesized a new image of the scene from different views of the input image. They then synthesized the target image with per-pixel depth information by adding pose and visibility in consecutive views. Inspired by this work, Wang et al. [47] utilized the camera parameters to synthesize the original view in a differential way.

Different from the way of synthesizing the target view, GAN (Generative Adversarial Network) is more straightforward. MonoGAN [25] first applied GAN to monocular depth estimation. Its network structure is similar to Fig.3(d). The intuition behind this is that the generator utilizes the image to obtain the depth map and then uses the depth map to warp the image. Next, the discriminator was used to classify the raw image as real and the warped image as fake. Compare with MonoGAN, the generator in [48] consists of DepthNet and PoseNet, which are used for depth map estimation and camera pose estimation, respectively, and uses the estimated depth map and camera pose parameters to generate image pairs.

Although using deep learning alone can achieve good results, it is still a problem how to bring geometry information or semantic information to depth estimation to improve the performance. Epipolar constraints play an important in geometry-based methods. Thus, epipolar constraints were used in learning-based methods to ensure proper pixel correspondence [49] and reduce the network parameters [50]. Its network structure is similar to Fig.3(b) Similar to geometry-based methods using stereo image sequence to obtain absolute scale depth, learning-based methods also used stereo image sequences to train the network to solve the scale ambiguity problems [51], [52]. To enforce consistency between the left and right disparities, bilateral cyclic consistency constraints were used in [53]. 2D photometric and 3D geometric information was also used in [54], which can consider the global scene and its geometric information. Besides, optical flow [55] and semantic information was also used into depth estimation problem [56].

3) Discussion : For VO or SLAM, obtaining a coarse depth information is a great improvement to the convergence and robustness. An overview of main approaches discussed in this section and whether they meet our criteria is described in table IV. Although stereo cameras can obtain depth information by utilizing stereo matching algorithms, it needs enormous computation. Monocular cameras have the characteristics of low cost and its application is becoming more popular. Therefore, obtaining depth information directly from monocular images has received great attention.

Based on table IV, we can see that almost all approaches can predict depth under outdoor environments, which is the basis for the system to work in complex and large-scale environments. However, most approaches cannot meet the criteria of efficiency because of deep model usually needs enormous computation, which cannot run real-time in resources limited systems. Further, it hard to apply to real-world applications. Besides, the supervised method relying solely on the dataset has poor generalization performance, making it difficult to obtain a model that is common to most scenarios. Therefore, unsupervised methods attract more attention.

From this short analysis, it seems that the problem of building an accurate, effective and practical algorithm or model remains a challenge. Regarding accuracy, even though some impressive results are shown, it is still difficult to deploy in real-world applications according to the aspect of practicality. Besides, it is also an important problem to explore what features the neural networks have learned in the problem of depth estimation.

TABLE IV
SUMMARY AND COMPARISON OF DEEP LEARNING TECHNIQUES FOR DEPTH ESTIMATION.

Ref	Method Type	Sensor	Environment Structure	Open Source	Ac	Ef	Sc	Dy	Pr	Ex
[35]	Supervised	Mo	In/Outdoor	Y	L	N	A	A	N	Y
[36]	Supervised	Mo	In/Outdoor	N	H	A	A	A	A	N
[34]	Supervised	Mo	In/Outdoor	N	H	N	A	A	N	N
[57]	Supervised	Mo	In/Outdoor	N	L	A	A	A	N	N
[31]	Supervised	Mo	In/Outdoor	N	H	A	A	A	A	N
[32]	Supervised	Mo	Indoor	Y	M	A	A	A	A	Y
[39]	Supervised	Mo	In/Outdoor	N	M	N	A	A	N	N
[41]	Supervised	Mo	In/Outdoor	Y	L	A	A	A	N	Y
[42]	Supervised	Mo	In/Outdoor	N	M	N	A	A	N	N
[58]	Supervised	St	Indoor	N	H	A	A	A	A	N
[44]	Unsupervised	Mo/St	Outdoor	N	M	A	A	A	N	N
[45]	Unsupervised	Mo/St	Outdoor	Y	H	A	A	A	Y	Y
[46]	Unsupervised	Mo	Outdoor	Y	M	Y	A	A	Y	Y
[59]	Unsupervised	Mo	Outdoor	N	L	A	A	A	N	N
[52]	Unsupervised	Mo/St	Outdoor	Y	M	A	A	A	Y	Y
[51]	Unsupervised	Mo	Outdoor	Y	L	Y	A	A	N	Y
[60]	Unsupervised	Mo	Outdoor	Y	H	N	A	A	Y	Y
[61]	Unsupervised	Mo	Outdoor	Y	M	Y	A	A	Y	Y
[25]	Unsupervised	Mo	Outdoor	Y	H	A	A	A	A	Y
[62]	Unsupervised	Mo	Outdoor	N	M	A	A	A	N	N
[56]	Unsupervised	Mo	Outdoor	Y	L	A	A	A	N	Y
[63]	Unsupervised	Mo	Outdoor	N	H	Y	A	A	Y	N
[54]	Unsupervised	Mo	Outdoor	Y	M	Y	A	A	Y	Y
[53]	Unsupervised	Mo	Outdoor	Y	M	A	A	A	A	Y
[64]	Unsupervised	Mo/St	Outdoor	N	H	N	A	A	N	N
[65]	Unsupervised	Mo/St	Outdoor	Y	H	A	A	A	Y	Y
[66]	Unsupervised	Mo	In/Outdoor	N	H	A	A	A	A	N
[67]	Unsupervised	Mo	Outdoor	N	H	A	A	A	A	N
[68]	Unsupervised	Mo	Outdoor	N	H	Y	A	Y	Y	N

Sensor: Mo (Monocular) **St** (Stereo)
Ac: Accuracy **Ef:** Efficiency **Sc:** Scalability **Dy:** Dynamicity **Pr:** Practicability **Ex:** Extensibility
Ac: H (High) M (Medial) L (Low) **Ef** & **Sc** & **Dy** & **Pr** & **Ex:** Y (Yes) N (No) A (Absence or No mention)

B. Deep learning for feature extraction and matching

1) Feature extraction : Feature points are usually composed of key points and descriptors. Finding and matching then across images has been attracted large attention of researchers. Fig.4 shows a comparison of different types of input, network architectures, loss functions, and output used to train the method of local image descriptor models. The descriptor can be used to capture important and distinctive pixel points in an image, which plays an important role in feature extraction and matching task. To learn consistent descriptors, a novel mixed-context loss and scale-aware sampling method were proposed in [69]. Mixed-context loss takes advantage of the scale consistency of Siamese loss and the fast learning ability of triplet loss. To learn compact binary descriptors, an unsupervised-based method was introduced in [70], which enforces the criterion of minimal loss quantization, uniformly distributed code and uncorrelated bits. Besides, Yi et al. [71] proposed a unified framework for local feature detection and descriptors, and Zeng et al. [72] used a 3D convolutional layer-based framework to learn local geometric descriptors.

From another perspective, previous work used deep layer's features, but shallower layer's features are more suitable for matching tasks. Therefore, HiLM [73] combined the different layers' features to learn more effective descriptors. Because deep learning needs numerous data in the network training process, it usually a time-costing process. Progressive sampling strategies [74], [75] were used to allow the network to traverse a large amount of training data in a short time. Besides, to improve the descriptor geometric invariance and discrimination, a subspace pooling [76] was proposed to instead of the max-pooling or average pooling.

Since we don't know which points are "interesting", we can't find true "interesting" points by using manually labeled data. Unsupervised methods can automatically find "interesting" points from the data. Thus, Savinov et al. [77] introduced a method that can generate repeated interesting points even if the image undergoes a transformation. Then, map the detected points to real-values and rank the points according to the real-values. Finally, the top/bottom quantiles of these points are used as interesting points. However, the scoring network response curve is constrained when the training process relies only on simple ranking losses. Therefore, a loss function was developed in [78] that can maximize the peak value of the response map and improve the repeatability of the learner.

2) Feature matching : For feature matching, most previous work focused on Siamese and Triplet loss functions. Gadot et al. [80] used Siamese CNN to learn the descriptors of both images, and then compared the learned descriptors using the L2 norm. The core of this method is a novel loss

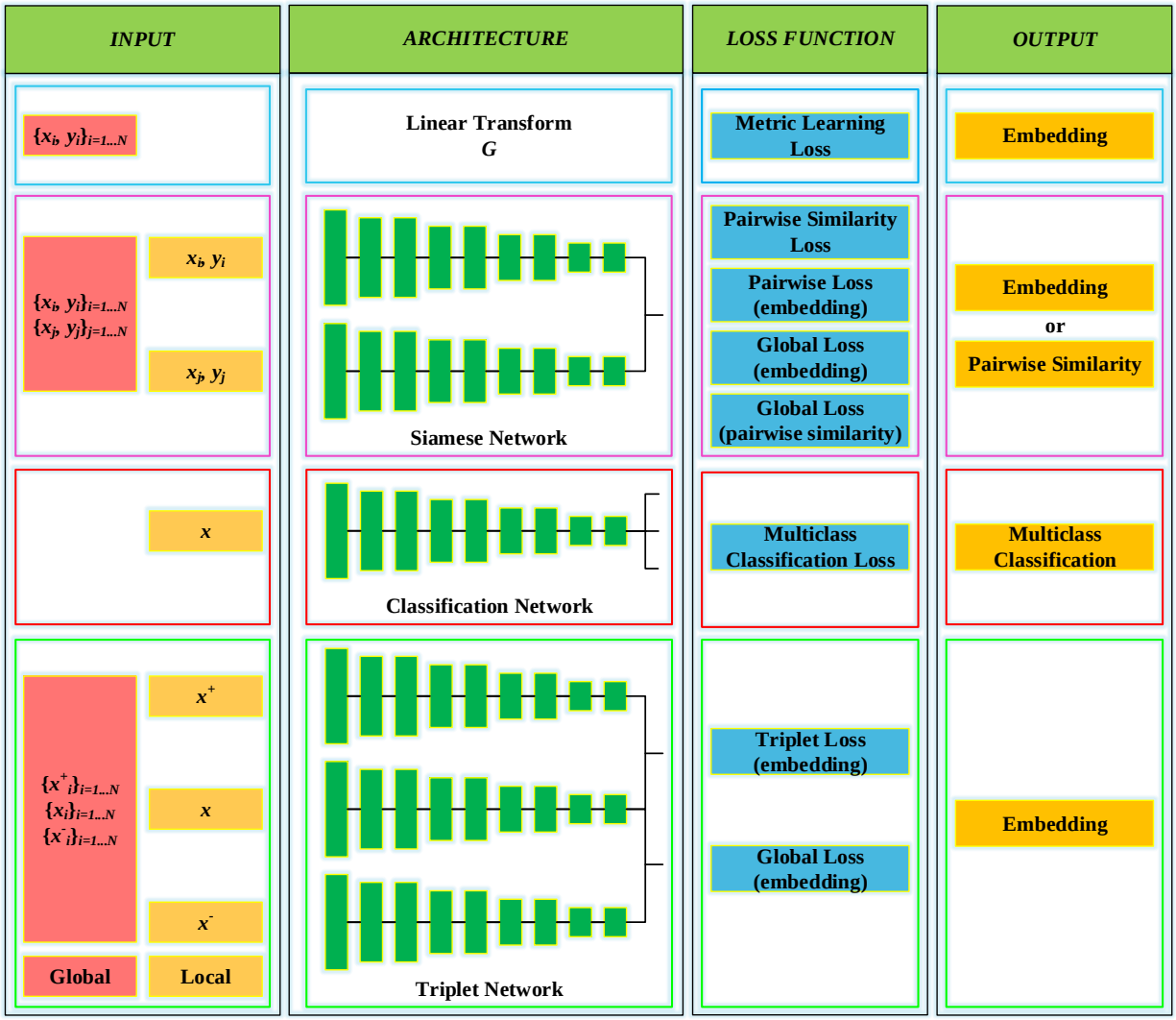

Fig. 4. Comparison of different types of input, network architectures, loss functions and output used to train the method of local image descriptor models [79]. From top to bottom: (a) Linear Transform G trained by the metric learning loss and produce a feature embedding. (b) Different types of loss functions and input types can be used to train the Siamese network and produce a feature embedding (x) or a pairwise similarity estimation (x_i, x_j). (c) The Classification Network can be used to classify local image descriptors and used for multiclass classification problems. (d) The Triplet Network can also be trained using different kinds of loss functions and input types and produces a feature embedding. x^+ represents a point from different class of x^+, x and x^- represents a point from a different class of x.

function, which can compute each training batch's higher loss distribution moments. Additionally, a global loss function was proposed in [79] that can minimize the mean distance between the same class descriptors and maximize the mean distance between the different class descriptors. Then, this work was extended in [81] by proposing a regularization term to spread out the local feature descriptor in descriptor space, and the proposed regularization can improve all methods using pairwise or triplet loss.

Contrast to traditional work that matches points according to the descriptors. Cieslewski et al. [82] proposed a method for matching interesting points without descriptors. The proposed network has multiple output channels so that the corresponding points of two images can be matched implicitly by the channel ids. For example, if the i-th interesting point is the maximum of the i-th channel, that point should match the i-th channel in another image. Because there are no matching descriptors, the system requires less memory and less computational cost. Although this method cannot achieve the performance of traditional methods, it can generate confidence for specific interesting points.

Further, a CNN- and RNN-based framework was proposed in [83]. The importance of this work is that the network does not learn some specific descriptors, but learns how to find the descriptor that can be matched well.

3) Discussion: Due to the proposed criteria is more focus on the full visual odometry system, we don't use these criteria to evaluate the literature about feature extraction and matching. But to a certain extent, learning-based methods can alleviate the algorithm's dependence on manual features. By utilizing the learning capability of deep learning, algorithms can directly learn more effective and robust feature representations.

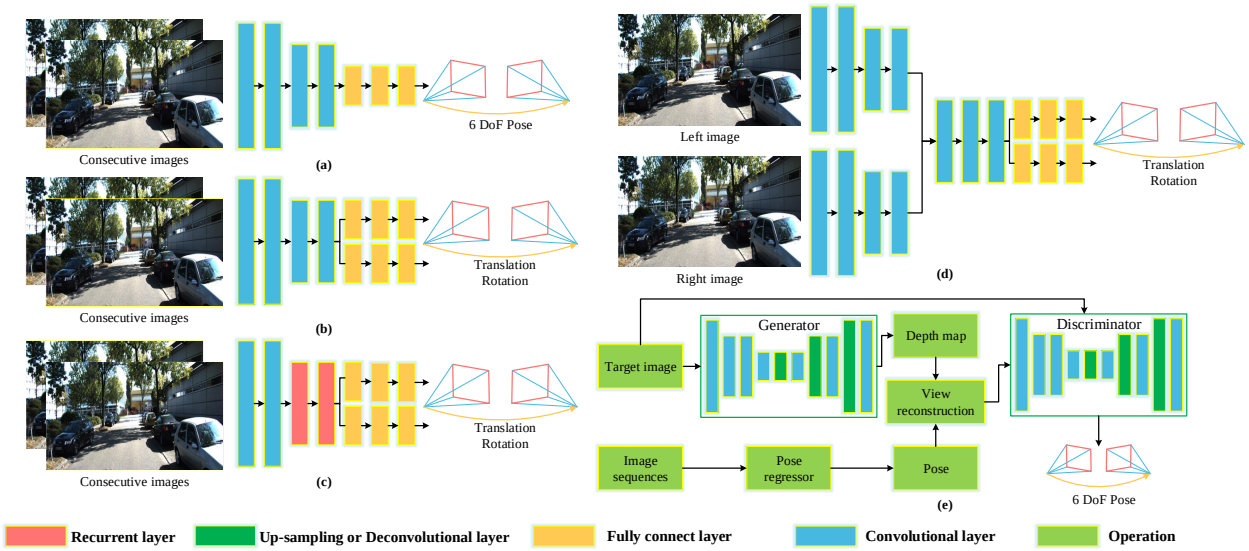

Fig. 5. Description of different neural network architectures for pose estimation. (a) CNN-based framework, which produces a 6 DoF pose. (b) CNN-based framework with two fully connected networks to produce translation and rotation, respectively. (c) RNN-based framework, which produces a forward-backward consistent pose estimation. (d) Stereo based framework, which uses stereo image to obtain absolute scale pose estimation. (e) GAN based framework.

However, the efficiency of learning-based methods still needs to be improved. With the popularity of mobile devices, fast and robust feature extraction and matching for mobile devices is also a trend for future development. On the other hand, the improvement of the accuracy of learning-based methods is based on large-scale data in the training process. Therefore, in some scenarios without sufficiency training dataset, it is still difficult to use deep learning techniques. Besides, is there something missing from the popular feature learning methods such as Siamese and Triplet architectures? Finally, although CNN can extract the information contained in the patch and establish the invariance of complex geometry and illumination changing, is this learned invariance overly rely on the data and cannot be effectively generalized to the complex changes between the images in the real world?

C. Deep learning for pose estimation

1) Supervised methods: PoseNet [84] was the work that the first to utilize CNN to regress 6 DOF camera pose from the monocular image in an end-to-end manner without graph optimization. Its network structure is similar to Fig.5(a). In large baselines that SIFT-based methods fail sharply, the proposed method can handle well. Besides, this method is also robust to illumination changing, motion blur and different camera intrinsics. Then, this work was extended in [85] by utilizing multi-view geometry as a source of training data to improve the PoseNet performance.

Since utilizing CNN to estimate camera pose from monocular images, it cannot model the relationship of consecutive images. Therefore, CNN and LSTM based pose regressor was introduced in [86], where CNN is used to learn suitable and robust feature representations, and LSTM is used to choose the most useful feature correlation for pose estimation. A spatial-temporal model based on bidirectional RNN was proposed in [87] to learn the relationship between consecutive images. MagicVO [88] used bidirectional LSTM to learn dynamic dependencies in consecutive images. In addition, DeepVO [24] used RNN to model the relationship of motion dynamic and image sequences. ESP-VO [23] then extended this work by adding uncertainty estimation for pose estimation. The RNN based framework is similar to Fig.5(c).

Due to the visual odometry will pass the error from the previous time to the next time, the error will accumulate over time, and drift is inevitable. To reduce the drift, an RCNN based framework DGRNet was proposed in [89]. In this framework, the pose estimation sub-network was used to smooth the visual odometry trajectory and the pose regression sub-network was used to reduce the accumulation of camera pose errors. Compared with this method using only image streams, Peretroukin et al. [90] used Bayesian Convolutional Neural Network to generate sun direction information in the image and then incorporated this sun direction information into stereo visual odometry pipeline to reduce drift.

For the environment that geometry-based methods hard to handle, learning-based methods show a strong capability. Costante et al. [91] proposed three different CNN architectures to learn robust representations to overcome the problems of blur, luminance and contrast anomalies. VLocNet [92] learn separate discriminative features that can be well generalized to motion blur and perceptual aliasing environments.

2) Unsupervised methods: For unsupervised learning-based methods, most of them use depth and optical flow to assist the training process [105]–[108]. DeMoN [97] was the first to use unsupervised learning methods to estimate both depth and pose from consecutive images. This network used optical flow to assist the depth and motion estimation. SfM-Net [106] simultaneously predict pixel-level depth, segmentation, and camera pose from consecutive images. D3VO [26] incorporates depth estimation, pose estimation and pose uncertainty into a framework to improve the performance. Further, to

TABLE V
SUMMARY AND COMPARISON OF DEEP LEARNING TECHNIQUES FOR POSE ESTIMATION.

Ref	Method Type	Sensor	Environment Structure	Open Source	Ac	Ef	Sc	Dy	Pr	Ex
[7]	Direct	Mo/St	In/Outdoor	Y	M	Y	Y	N	Y	Y
[5]	Feature	Mo/St/In	In/Outdoor	Y	H	Y	Y	N	Y	Y
[9]	Semi-Direct	Mo/St	In/Outdoor	Y	M	Y	Y	N	Y	Y
[84]	Supervised	Mo	In/Outdoor	Y	L	Y	Y	N	N	Y
[85]	Supervised	Mo	In/Outdoor	N	M	Y	Y	N	N	N
[86]	Supervised	Mo	In/Outdoor	N	M	A	Y	N	N	N
[87]	Supervised	Mo	In/Outdoor	N	M	Y	Y	N	Y	N
[88]	Supervised	Mo	In/Outdoor	N	H	A	Y	N	N	N
[24]	Supervised	Mo	Outdoor	Y	L	N	Y	N	N	Y
[23]	Supervised	Mo	In/Outdoor	N	L	N	Y	N	N	N
[91]	Supervised	Mo	Outdoor	N	M	Y	Y	N	Y	N
[93]	Supervised	St	Outdoor	Y	M	A	Y	N	N	Y
[92]	Supervised	Mo	Outdoor	N	H	Y	Y	N	Y	N
[94]	Supervised	Mo	Outdoor	N	M	A	Y	N	N	N
[95]	Supervised	Mo	Outdoor	N	H	Y	Y	N	Y	N
[96]	Supervised	St	In/Outdoor	N	H	N	Y	Y	Y	N
[97]	Unsupervised	Mo	In/Outdoor	Y	M	A	Y	N	N	Y
[54]	Unsupervised	Mo	Outdoor	Y	L	Y	Y	N	N	Y
[98]	Unsupervised	Mo	Outdoor	N	L	A	Y	N	N	N
[62]	Unsupervised	Mo	Outdoor	N	M	A	Y	N	N	N
[99]	Unsupervised	Mo	Outdoor	N	M	Y	Y	Y	Y	N
[61]	Unsupervised	Mo	Outdoor	Y	M	A	Y	Y	N	Y
[100]	Unsupervised	Mo	Indoor	Y	M	Y	N	N	N	Y
[52]	Unsupervised	Mo/St	Outdoor	Y	M	A	Y	N	N	Y
[51]	Unsupervised	Mo/St	Outdoor	Y	H	Y	Y	N	Y	Y
[60]	Unsupervised	Mo/St	Outdoor	Y	M	N	Y	N	N	Y
[101]	Unsupervised	Mo	Outdoor	N	H	N	Y	N	N	N
[102]	Unsupervised	Mo	Outdoor	Y	H	A	Y	Y	N	Y
[103]	Unsupervised	Mo	Outdoor	N	M	A	Y	N	N	N
[104]	Unsupervised	Mo	In/Outdoor	N	H	Y	Y	N	Y	Y

Sensor: Mo (Monocular) **St** (Stereo) **In** (Inertial)
Ac: Accuracy **Ef:** Efficiency **Sc:** Scalability **Dy:** Dynamicity **Pr:** Practicability **Ex:** Extensibility
Ac: H (High) M (Medial) L (Low) **Ef** & **Sc** & **Dy** & **Pr** & **Ex:** Y (Yes) N (No) A (Absence or No mention)

learn consistent 3D structures and better exploit unsupervision, a forward-backward constraint based on the left-right consistency constraint [45] was proposed. In contrast, Iyer et al. [100] proposed Composite Transformation Constraints to generate supervisory signals during training and to enforce geometric consistency. Besides, Almalioglu et al. [62] use generative adversarial networks (GANs) for camera pose estimation from monocular image sequences. Its network structure is similar to Fig.5(e).

The monocular visual odometry faced the scale ambiguity problem. Many methods used stereo images to train networks to remove this problem [51], [52], [60]. Its network structure is similar to Fig.5(d). Besides, different from using stereo images to obtain absolute scale estimation, using depth information obtained from 3D LIDARs to train the network to obtain absolute scale estimation was introduced in [101].

Another problem that plagues VO is how to alleviate the impact of dynamic objects. To reduce the distraction of dynamic objects in the scene, Barnes et al. [99] introduced an ephemerality mask that can estimate the likelihood that pixels in the input image correspond to static or dynamic objects in the scene. Additionally, GeoNet was proposed in [61], which consists of two stages and three sub-networks. The first stage is a rigid structure re-constructor for depth and pose estimation, and the last stage is a non-rigid motion localizer for optical flow estimation. Specifically, three subnetworks extract geometric relationships separately, and then combine as an image reconstruction loss to infer the static and dynamic scene parts in two stages, respectively.

Since geometry-based methods becoming mature, how to utilize this geometry knowledge to improve learning-based methods is also a problem. To consider ambiguous pixels and obtain better geometric understanding ability, Prasad et al. [49] proposed an epipolar geometry-based approach, which gives each pixel a weight according to whether it is projected correctly. Its network structure is similar to Fig.5(b). Shen et al. [102] proposed to incorporate intermediate geometric information such as pairwise matching to the pose estimation problem to solve the problem of results involving large systematic errors under realistic scenarios. VLocNet++ [103] embedded geometric and semantic knowledge of the scene into the pose regression network.

3) Discussion: An overview of the main approaches discussed in this section and how they whether meets our criteria is described in table V. As shown in table V, the current learning-based methods still not yet viable for robots working under dynamic environments. It remains a challenge for both learning-based and geometry-based methods. Another

interesting observation is that even the accuracy of learning-based methods is high that they also cannot be used in practice. The reason is that the learning-based model inference is a time-consuming process. Of course, things could evolve in the near future depending on the system computation power increasing. However, designing a lighter and small network with good results is also an evolving trend.

For accuracy, some currently existing learning-based methods can compare with the state-of-the-art geometry-based methods. However, it builds on the similarity (feature representation) between the training and testing dataset. If testing environments contain many 'unseen' scenarios, the performance will be decreased. However, a dataset contains all scenarios is impossible. For efficiency, most applications require to produce online estimation in a timely fashion. Unfortunately, they are resources limited and do not have GPU to accelerate inference. Last but not least, robustness. Due to the powerful learning capability of deep learning, it can improve the robustness to a certain extent. Nevertheless, this robustness is heavily dependent on large-scale data. We don't know what the algorithm learned in this process. How to use existing excellent geometry-based algorithms to guide them to improve the robustness in challenging environments is a promising way.

IV. Application of Visual Odometry in Complicated and Emerging Areas

In this section, we will focus on visual odometry applications. Recently, visual odometry has a wide range of applications and has been effectively applied in several fields, as shown in Fig.6.

In mobile robotic systems, the autonomous vehicle is a rapidly advancing application area of visual odometry with unstructured environments [109]–[111]. Visual odometry is generally regarded as a key for truly autonomous vehicles. In this application, the algorithm more focus on dynamic [99], illumination changing [112], long-term [113] and large-scale environments [114]. Compared with autonomous vehicles, using visual odometry into UAVs is more challenges due to the computational capability and 3D maneuverability [115], [116]. In order to obtain more robust performance, many researches have tried to fusing data from multi-modal sensors, such as IMU [117], [118]. Recently, there has shown a growing interest in fusing multiple cameras [119]–[121]. Besides, visual odometry is also widely used in underwater robots [122]–[124], space exploration robots [125]–[127] and agriculture robots [128], etc.

In the medical industry, VO has great potential in image-based medical applications, which can navigate outside or inside the patient's body to refer position and discovered problems [129]. For example, active wireless capsule endoscopes used VO techniques to track his location without additional sensors and hardware in the GI tract [130]–[133]. Besides, in surgical applications, visual odometry also gained a big attention such as surgical systems, surgical assistance robot and image-guided surgery systems [134]–[136].

In augmented reality (AR), VO also plays an important role. Visual odometry are used to obtain the device real world

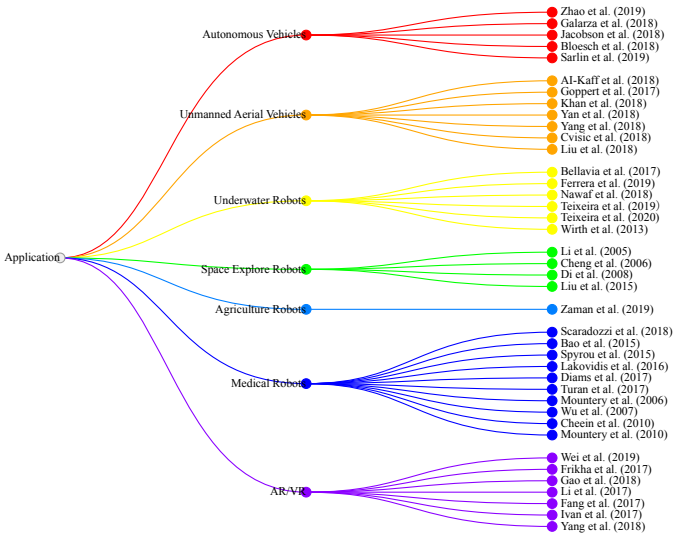

Fig. 6. Application of visual odometry.

coordinates independent of camera and camera images [137]. Then, integrate augmentations with the real world according to real world coordinates obtained by VO. In the beginning, these systems usually use camera as only sensors [138], [139]. In mobile devices, to improve accuracy and obtain absolute scale, the devices fuse data from other sensors such as IMU [140]–[143].

Although visual odometry has a wide range of potential applications, the different application requires special hardware and software design. Each of these applications brings different challenges. It seems that deep learning is a promising way to handle these challenges. It has been demonstrated that deep learning can improve accuracy and robustness. Deep VO will play a big role in complicated and rising areas, including search and rescue, planetary exploration, service robots.

V. Challenges and Opportunities

A. Challenges

First, the data available to train the deep learning model for various tasks is insufficient. Although we know that large-scale high-quality data is required to train deep models, there is still a limited number of the large-scale dataset available in visual odometry or SLAM. The existing large-scale dataset like KITTI from the perspective of geometry-based methods is not large-scale from the perspective of deep learning. Besides, the scene of these existing datasets is single, which causes poor generalization ability of the model. The performance of deep learning-based methods can be improved as the amounts of datasets increase. Therefore, a large-scale and high-quality dataset is important to push learning-based methods, like ImageNet for object recognition and PASCAL VOC dataset for object detection.

Second, deep learning methods used for visual odometry are simple. Since CNN surpasses human ability to classify object recognition in ImageNet 1000 challenge, CNN has almost dominated the field of computer vision. Therefore, CNN was

first used to estimate the camera pose and has spawned a lot of work based on CNN. However, CNN is inadequate and less effective to learn relationships and dynamics from sequential data. Thus, RNN was used to learn these complex connections and dependencies in an end-to-end manner. Different methods have their own disadvantages and advantages. A problem of RNN is that they cannot extract effective feature representation from high-dimensional data like CNN. The power of deep learning should be far more than this. Therefore, bring new techniques and theories to visual odometry needs us to explore in the future.

Third, we don't know what features the neural networks have learned. For geometry-based methods, it is explicit and well understand. Although it is not robust to dynamic and illumination changing environments and needs plenty of time to manually fine-tune, we can know what the algorithm learned from this process. Nevertheless, for deep learning, we think more of it as a black box. We usually can only rely on the trained model to obtain the results and cannot predict theoretically. Even if the model results not meet our expectations, we do not know how to improve.

Fourth, we lack a unified and comprehensive benchmark to evaluate existing algorithms. Currently, we usually use RMSE or AE to evaluate algorithms accuracy. However, we often encounter a problem that is why when we deploy the algorithm in the real world, its performance cannot reach the published performance in the paper? The reason is that the real world performance affected by various factors not only accuracy but also robustness, generalization, etc. However, we are currently unable to quantitatively evaluate robustness, generalization, etc. In this paper, we only proposed some criteria to evaluate in qualitative. However, we need a benchmark to evaluate in quantitative, not only focusing on statistic (EMSE, ATE) over the dataset, but also using the human behavior across various conditions as the reference. In this way, we can compare the gaps between different models and the potential gaps between humans and models to improve our algorithms and select the algorithm for specific problem. However, how to quantitative these criteria remains a challenge.

Finally, the generalization ability and robustness of current algorithms need to be improved. Compared with geometry-based methods, learning-based algorithms usually don't fail initialize and lose track. The reason is that given an image, the model always produces a prediction. However, it can make a big error due to wrong or "unseen" input. Besides, the similarity of feature representations between the training and testing dataset are the decisive factors affecting performance. Therefore, how to improve the generalization is a big problem. Based on table III, few literatures discussed the dynamic environments. Although the capability of deep learning improved the accuracy and robustness in specific situations, learning-based methods are still lacking an effective way to solve this problem. To improve the robustness of the algorithm is an important step to robust perception.

B. Opportunities

(1) Use unsupervised learning techniques to train the network. Since deep learning models trained by supervised learning acquiring amount of human-labeled data, it cannot benefit from large-scale unlabeled data. The performance of unsupervised learning can be improved as the amounts of datasets increase. On the one hand, unsupervised learning can truly exploit high-dimension features from large-scale data. On the other hand, we can use the geometry-based loss function to guide the learning process in unsupervised learning.

(2) Use semantic information to obtain semantic reasoning ability. Understanding semantic information is the most significant step to obtain high-level understanding and semantic reasoning. Currently, robots only understanding the low-level geometric features but they don't truly exploit semantics. On the one hand, utilizing semantic information and object detection results can form semantic-level based localization constraints, thereby improving the accuracy and robustness. On the other hand, integrating semantic information into the visual odometry allows the robot to infer the surrounding environment. For example, when the system detects a vehicle, the system can infer the direction of the vehicle's motion based on the direction and position of the vehicle. Moreover, understanding high-level semantic information, object properties and the properties' mutual relations can provide better interaction between the robot and the environment.

(3) Try to fuse data from multi-modal sensors by using deep learning techniques. Since a single sensor is difficult to meet the requirement, many systems combined camera data with multi-modal sensor data by using filters or optimization techniques. However, few literatures focus on learning-based fusion methods [144]–[147]. Compared with traditional fusion methods, learning-based methods do not require manual complex system modeling and calibration, such as synchronization and calibration between the camera and other sensors, modeling the sensor's noise and biases, removing gravity from acceleration measurements according to orientation. Therefore, learning-based multi-sensor fusion methods have the potential to generate new adaptive VO paradigms that can be adapted to different sensors.

(4) Combining deep learning-based methods with geometry-based methods. Geometry-based methods are more reliable and accurate under static, sufficiency illumination and texture environments. However, they still face many problems in challenging environments. While learning-based methods can handle challenging environments, these methods cannot provide reliability and accuracy of geometry-based methods in conditions that suitable for geometry-based methods. Therefore, combining geometry-based methods with deep learning-based methods is an effective approach to improve system accuracy and robustness. However, it is not clear in which way this should be realized to be most effective.

(5) Deploy learning-based algorithms in hardware platform efficiently. Currently, most state-of-the-art algorithms with tremendous parameters require a high-end GPUs to run inference in real-time. However, robotic platforms usually only have CPUs or low-powered GPUs, which cannot affordable for some learning-based algorithms to operate in real-time. Therefore, there are many restrictions to deploy these algorithms to embedded platforms. The size of the network structure is the main influencing factor. Generally, large networks need more

computing power and memory for reasoning, which leads to low efficiency. Besides, the size of the network also affects accuracy. Therefore, designing the lighter, smaller network with good enough results is a significant step to make these algorithms into practice.

(6) Use advanced structures and techniques of deep learning in visual odometry. After the great development of deep learning in recent years, it is gratifying to see a wide range of deep learning techniques. However, now we only use some simple and general techniques, such as CNN, RNN, GAN, and its hybrids [23], [24], [37], [38], [86], [87]. Therefore, we should explore more possibilities to use advanced deep learning techniques. For example, to avoid drift and accumulated errors, we usually need build a large-scale optimization problem by using graph-based optimization methods. Therefore, we can take GNN (Graph Neural Network) into the existing work to obtain the optimal trajectory. Besides, attention-based network can make this process more human-like. In [148], a spatial-temporal attention mechanism is applied to conduct the contexts for feature selection. In [149], they use attention mechanism dynamically adjusts the semantic categories weights to alleviate the influence of dynamic objects.

VI. Conclusion

The problem of deep learning-based visual odometry has achieved great progress over the past five years. Through reviewing these significant works, we can see that several important issues have been solved, many challenging problems have been raised, and some new applications also have been developed. We have witnessed the ongoing evolution of visual odometry from the classical age to the weak artificial intelligence age, and finally to the strong artificial intelligence age. The classic age refers to use of visual geometric, probabilistic, and filtering approaches to solve this problem. The weak artificial intelligence age refers to use of machine learning and deep learning approaches to build solutions, which is also the main focus of this paper. Now, we are in the stage of evolution of the weak artificial intelligence age to the strong artificial intelligence age. As we saying beginning, in the strong artificial intelligence age, it becomes possible for the robot to acquire a high-level understanding of the environments and achieve active (task-driven) perception. By combining visual geometry, deep learning comparing with the knowledge of human cognition and development, the VO system can provide insightful, intuitive, and compact environment models with the one created and used by humans. In this process, visual psychophysics will play an important role by integrating the inner mechanisms of human visual processing with deep learning. VO still constitutes an indispensable module for most robotics, such as cognitive and development robot, autonomous vehicle, etc. Therefore, to make VO system more accurate, robust and insightful, we still have many interesting but challenging problems to solve.

Acknowledgment

The authors thank the financial support of National Natural Science Foundation of China (Grant No: 51605054), Key Technical Innovation Projects of Chongqing Artificial Intelligent Technology (Grant No. cstc2017rgzn-zdyfX0039), Chongqing Social Science Planning Project (No:2018QNJJ16), Fundamental Research Funds for the Central Universities (No: 2019CDXYQC003).

References

[1] Y. Q. Ma, S. C. Liu, B. Sima, B. Wen, S. Peng, and Y. Jia, "A precise visual localisation method for the chinese chang'e-4 yutu-2 rover," *Photogrammetric Record*, vol. 35, no. 169, pp. 10–39, 2020.

[2] H. Tang, R. Yan, and K. C. Tan, "Cognitive navigation by neuro-inspired localization, mapping, and episodic memory," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 751–761, 2018.

[3] A. Taniguchi, T. Taniguchi, and T. Inamura, "Spatial concept acquisition for a mobile robot that integrates self-localization and unsupervised word discovery from spoken sentences," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 4, pp. 285–297, 2016.

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: A versatile and accurate monocular slam system," *Ieee Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[5] R. R. Campos, C.and Elvira, J. J. G. Montiel, J. M., Tardos, and J. D., "Orb-slam3: An accurate open-source library for visual, visual-inertial and multi-map slam," *arXiv preprint*, vol. arXiv:2007.11898, 2020.

[6] A. Vakhitov, V. Lempitsky, and Y. Zheng, "Stereo relative pose from line and point feature triplets," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Conference Proceedings, pp. 648–663.

[7] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Trans Pattern Anal Mach Intell*, vol. 40, no. 3, pp. 611–625, 2018.

[8] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*. Springer, Conference Proceedings, pp. 834–849.

[9] J. Engel, J. Sturm, and D. Cremers, "Semi-dense visual odometry for a monocular camera," in *Proceedings of the IEEE international conference on computer vision*, Conference Proceedings, pp. 1449–1456.

[10] M. A. Esfahani, H. Wang, K. Y. Wu, and S. H. Yuan, "Aboldeepio: A novel deep inertial odometry network for autonomous vehicles," *Ieee Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 1941–1950, 2020.

[11] F. W. Ma, J. Z. Shi, L. H. Ge, K. Dai, S. R. Zhong, and L. Wu, "Progress in research on monocular visual odometry of autonomous vehicles," *Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)*, vol. 50, no. 3, pp. 765–775, 2020.

[12] J. C. Piao and S. D. Kim, "Real-time visual-inertial slam based on adaptive keyframe selection for mobile ar applications," *Ieee Transactions on Multimedia*, vol. 21, no. 11, pp. 2827–2836, 2019.

[13] B. Talbot, F. Dayoub, P. Corke, and G. Wyeth, "Robot navigation in unseen spaces using an abstract map," *IEEE Transactions on Cognitive and Developmental Systems*, pp. 1–1, 2020.

[14] Y. Wang, L. Zhang, L. Wang, and Z. Wang, "Multitask learning for object localization with deep reinforcement learning," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, no. 4, pp. 573–580, 2019.

[15] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendon-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 55–81, 2015.

[16] M. O. Aqel, M. H. Marhaban, M. I. Saripan, and N. B. Ismail, "Review of visual odometry: types, approaches, challenges, and applications," *Springerplus*, vol. 5, no. 1, p. 1897, 2016.

[17] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *Ieee Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, 2016.

[18] S. D. Huang and G. Dissanayake, "A critique of current developments in simultaneous localization and mapping," *International Journal of Advanced Robotic Systems*, vol. 13, 2016.

[19] S. Saeedi, M. Trentini, M. Seto, and H. Li, "Multiple-robot simultaneous localization and mapping: A review," *Journal of Field Robotics*, vol. 33, no. 1, pp. 3–46, 2016.

[20] R. H. Li, S. Wang, and D. B. Gu, "Ongoing evolution of visual slam from geometry to deep learning: Challenges and opportunities," *Cognitive Computation*, vol. 10, no. 6, pp. 875–889, 2018.

[21] M. R. U. Saputra, A. Markham, and N. Trigoni, "Visual slam and structure from motion in dynamic environments: A survey," *Acm Computing Surveys*, vol. 51, no. 2, 2018.

[22] M. Sualeh and G. W. Kim, "Simultaneous localization and mapping in the epoch of semantics: A survey," *International Journal of Control Automation and Systems*, vol. 17, no. 3, pp. 729–742, 2019.

[23] S. Wang, R. Clark, H. K. Wen, and N. Trigoni, "End-to-end, sequence-to-sequence probabilistic visual odometry through deep neural networks," *International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 513–542, 2018.

[24] S. Wang, R. Clark, H. Wen, and N. Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, Conference Proceedings, pp. 2043–2050.

[25] P. M. Aleotti F, Tosi F, "Generative adversarial networks for unsupervised monocular depth prediction," *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[26] N. Yang, L. v. Stumberg, R. Wang, and D. Cremers, "D3vo: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Conference Proceedings, pp. 1281–1292.

[27] U. R. Geiger A, Lenz P, "Are we ready for autonomous driving? the kitti vision benchmark suite," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[28] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Information Processing Systems 27 (Nips 2014)*, vol. 27, 2014.

[29] S. Bazrafkan, H. Javidnia, J. Lemley, and P. Corcoran, "Semiparallel deep neural network hybrid architecture: first application on depth from monocular camera," *Journal of Electronic Imaging*, vol. 27, no. 4, 2018.

[30] Y. Kim, H. Jung, D. Min, and K. Sohn, "Deep monocular depth estimation via integration of global and local predictions," *Ieee Transactions on Image Processing*, vol. 27, no. 8, pp. 4131–4144, 2018.

[31] B. Li, Y. C. Dai, and M. Y. He, "Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference," *Pattern Recognition*, vol. 83, pp. 328–339, 2018.

[32] D. Xu, E. Ricci, W. L. Ouyang, X. G. Wang, and N. Sebe, "Multi-scale continuous crfs as sequential deep networks for monocular depth estimation," *30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017)*, pp. 161–169, 2017.

[33] Y. F. Xu, Y. Wang, and L. Guo, "Unsupervised ego-motion and dense depth estimation with monocular video," *2018 Ieee 18th International Conference on Communication Technology (Icct)*, pp. 1306–1310, 2018.

[34] Z. Y. Zhang, C. Y. Xu, J. Yang, Y. Tai, and L. Chen, "Deep hierarchical guidance and regularization learning for end-to-end depth estimation," *Pattern Recognition*, vol. 83, pp. 430–442, 2018.

[35] F. Y. Liu, C. H. Shen, G. S. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *Ieee Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2024–2039, 2016.

[36] Y. Z. H. Cao, Z. F. Wu, and C. H. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *Ieee Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3174–3182, 2018.

[37] A. Grigorev, F. Jiang, S. Rho, W. J. Sori, S. H. Liu, and S. Sai, "Depth estimation from single monocular images using deep hybrid network," *Multimedia Tools and Applications*, vol. 76, no. 18, pp. 18 585–18 604, 2017.

[38] A. C. S. Kumar, S. M. Bhandarkar, and M. Prasad, "Depthnet: A recurrent neural network architecture for monocular depth prediction," *Proceedings 2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition Workshops (Cvprw)*, pp. 396–404, 2018.

[39] H. Yan, S. L. Zhang, Y. Zhang, and L. Zhang, "Monocular depth estimation with guidance of surface normal map," *Neurocomputing*, vol. 280, pp. 86–100, 2018.

[40] L. He, G. Wang, and Z. Hu, "Learning depth from single images with deep neural network embedding focal length," *IEEE Trans Image Process*, vol. 27, no. 9, pp. 4676–4689, 2018.

[41] F. C. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," *2018 Ieee International Conference on Robotics and Automation (Icra)*, pp. 4796–4803, 2018.

[42] Z. Zhang, C. Xu, J. Yang, J. Gao, and Z. Cui, "Progressive hard-mining network for monocular depth estimation," *IEEE Trans Image Process*, vol. 27, no. 8, pp. 3691–3702, 2018.

[43] M. I. D. Zhang Y, Nguyen T, "Dfinenet: Ego-motion estimation and depth refinement from sparse, noisy depth input with rgb guidance," *arXiv preprint arXiv:1903.06397*, 2019.

[44] R. Garg, B. G. VijayKumar, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," *Computer Vision - Eccv 2016, Pt Viii*, vol. 9912, pp. 740–756, 2016.

[45] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," *30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017)*, pp. 6602–6611, 2017.

[46] T. H. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," *30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017)*, pp. 6612–+, 2017.

[47] A. J. Wang, Z. J. Fang, Y. B. Gao, X. Y. Jiang, and S. W. Ma, "Depth estimation of video sequences with perceptual losses," *Ieee Access*, vol. 6, pp. 30 536–30 546, 2018.

[48] A. C. S. Kumar, S. M. Bhandarkar, and M. Prasad, "Monocular depth prediction using generative adversarial networks," *Proceedings 2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition Workshops (Cvprw)*, pp. 413–421, 2018.

[49] B. B. Prasad V, Das D, "Epipolar geometry based learning of multi-view depth and ego-motion from monocular sequences," *arXiv preprint arXiv:1812.11922*, 2018.

[50] V. Prasad and B. Bhowmick, "Sfmlearner++: Learning monocular depth and ego-motion using meaningful geometric constraints," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Conference Proceedings, pp. 2087–2096.

[51] R. H. Li, S. Wang, Z. Q. Long, and D. B. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," *2018 Ieee International Conference on Robotics and Automation (Icra)*, pp. 7286–7291, 2018.

[52] H. Y. Zhan, R. Garg, C. S. Weerasekera, K. J. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," *2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr)*, pp. 340–349, 2018.

[53] S. S. Wong A, Hong B W, "Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction," *arXiv preprint arXiv:1903.07309*, 2019.

[54] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," *2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr)*, pp. 5667–5675, 2018.

[55] Q. R. Teng, Y. M. Chen, and C. Huang, "Occlusion-aware unsupervised learning of monocular depth, optical flow and camera pose with geometric constraints," *Future Internet*, vol. 10, no. 10, p. 92, 2018.

[56] T. F. Ramirez P Z, Poggi M, "Geometry meets semantics for semi-supervised monocular depth estimation," *arXiv preprint arXiv:1810.04093*, 2018.

[57] Y. Kim, H. Jung, D. Min, and K. Sohn, "Deep monocular depth estimation via integration of global and local predictions," *IEEE Trans Image Process*, vol. 27, no. 8, pp. 4131–4144, 2018.

[58] U. Kusupati, S. Cheng, R. Chen, and H. Su, "Normal assisted stereo depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Conference Proceedings, pp. 2189–2199.

[59] V. Prasad, D. Das, and B. Bhowmick, "Epipolar geometry based learning of multi-view depth and ego-motion from monocular sequences," *arXiv preprint arXiv:.11922*, 2018.

[60] V. M. Babu, K. Das, A. Majumdar, and S. Kumar, "Undemon: Unsupervised deep network for depth and ego-motion estimation," *2018 Ieee/Rsj International Conference on Intelligent Robots and Systems (Iros)*, pp. 1082–1088, 2018.

[61] Z. C. Yin and J. P. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," *2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr)*, pp. 1983–1992, 2018.

[62] Y. Almalioglu, M. R. U. Saputra, P. P. de Gusmao, A. Markham, and N. Trigoni, "Ganvo: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks," 2018.

[63] S. N. Pilzer A, Lathuilière S, "Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation," *arXiv preprint arXiv:1903.04202*, 2019.

[64] P. M. Tosi F, Aleotti F, "Learning monocular depth estimation infusing traditional stereo knowledge," *arXiv preprint arXiv:1904.04144*, 2019.

[65] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "On the uncertainty of self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Conference Proceedings, pp. 3227–3237.

[66] A. Wang, Z. Fang, Y. Gao, S. Tan, S. Wang, S. Ma, and J. N. Hwang, "Adversarial learning for joint optimization of depth and ego-motion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4130–4142, 2020.

[67] A. Johnston and G. Carneiro, "Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Conference Proceedings, pp. 4756–4765.

[68] L. P. Zhou and M. Kaess, "Windowed bundle adjustment framework for unsupervised learning of monocular depth estimation with u-net extension and clip loss," *Ieee Robotics and Automation Letters*, vol. 5, no. 2, pp. 3283–3290, 2020.

[69] M. Keller, Z. T. Chen, F. Maffra, P. Schmuck, and M. Chli, "Learning deep descriptors with scale-aware triplet networks," *2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr)*, pp. 2762–2770, 2018.

[70] K. V. Lin, J. W. Lu, C. S. Chen, and J. Zhou, "Learning compact binary descriptors with unsupervised deep neural networks," *2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr)*, pp. 1183–1192, 2016.

[71] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "Lift: Learned invariant feature transform," *Computer Vision - Eccv 2016, Pt Vi*, vol. 9910, pp. 467–483, 2016.

[72] A. Zeng, S. R. Song, M. Niessner, M. Fisher, J. X. Xiao, and T. Funkhouser, "3dmatch: Learning local geometric descriptors from rgb-d reconstructions," *30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017)*, pp. 199–208, 2017.

[73] Z. Z. M. Fathy M E, Tran Q H, "Hierarchical metric learning and matching for 2d and 3d geometric correspondences," *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[74] A. Mishchuk, D. Mishkin, F. Radenovic, and J. Matas, "Working hard to know your neighbor's margins: Local descriptor learning loss," *Advances in Neural Information Processing Systems 30 (Nips 2017)*, vol. 30, 2017.

[75] Y. R. Tian, B. Fan, and F. C. Wu, "L2-net: Deep learning of discriminative patch descriptor in euclidean space," *30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017)*, pp. 6128–6136, 2017.

[76] X. Wei, Y. Zhang, Y. H. Gong, and N. N. Zheng, "Kernelized subspace pooling for deep local descriptors," *2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr)*, pp. 1867–1875, 2018.

[77] N. Savinov, A. Seki, L. Ladicky, T. Sattler, and M. Pollefeys, "Quad-networks: unsupervised learning to rank for interest point detection," *30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017)*, pp. 3929–3937, 2017.

[78] L. G. Zhang and S. Rusinkiewicz, "Learning to detect features in texture images," *2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr)*, pp. 6325–6333, 2018.

[79] B. G. V. Kumar, G. Carneiro, and I. Reid, "Learning local image descriptors with deep siamese and triplet convolutional networks by minimizing global loss functions," *2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr)*, pp. 5385–5394, 2016.

[80] D. Gadot and L. Wolf, "Patchbatch: a batch augmented loss for optical flow," *2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr)*, pp. 4236–4245, 2016.

[81] X. Zhang, F. X. Yu, S. Kumar, and S. F. Chang, "Learning spread-out local feature descriptors," *2017 Ieee International Conference on Computer Vision (Iccv)*, pp. 4605–4613, 2017.

[82] S. D. Cieslewski T, Bloesch M, "Matching features without descriptors: Implicitly matched interest points (imips)," *arXiv preprint*, vol. arXiv:1811.10681, 2018.

[83] J. P. Tang J, Folkesson J, "Geometric correspondence network for camera motion estimation," *IEEE Robotics and Automation Letters*, 2018.

[84] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," *2015 Ieee International Conference on Computer Vision (Iccv)*, pp. 2938–2946, 2015.

[85] A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," *30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017)*, pp. 6555–6564, 2017.

[86] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using lstms for structured feature correlation," *2017 Ieee International Conference on Computer Vision (Iccv)*, pp. 627–637, 2017.

[87] R. Clark, S. Wang, A. Markham, N. Trigoni, and H. K. Wen, "Vidloc: A deep spatio-temporal model for 6-dof video-clip relocalization," *30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017)*, pp. 2652–2660, 2017.

[88] J. Jiao, J. Jiao, Y. Mo, W. Liu, and Z. J. a. p. a. Deng, "Magicvo: End-to-end monocular visual odometry through deep bi-directional recurrent convolutional neural network," 2018.

[89] Y. Lin, Z. Liu, J. Huang, C. Wang, G. Du, J. Bai, S. Lian, and B. Huang, "Deep global-relative networks for end-to-end 6-dof visual localization and odometry," 2018.

[90] V. Peretroukhin, L. Clement, and J. Kelly, "Inferring sun direction to improve visual odometry: A deep learning approach," *International Journal of Robotics Research*, vol. 37, no. 9, pp. 996–1016, 2018.

[91] G. Costante, M. Mancini, P. Valigi, T. A. J. I. r. Ciarfuglia, and a. letters, "Exploring representation learning with cnns for frame-to-frame ego-motion estimation," *IEEE ROBOTICS AND AUTOMATION LETTERS*, vol. 1, no. 1, pp. 18–25, 2015.

[92] A. Valada, N. Radwan, and W. Burgard, "Deep auxiliary learning for visual localization and odometry," *2018 Ieee International Conference on Robotics and Automation (Icra)*, pp. 6939–6946, 2018.

[93] V. Peretroukhin and J. Kelly, *IEEE Robotics and Letters*.

[94] B. B. Prasad V, "Sfmlearner++: Learning monocular depth and ego-motion using meaningful geometric constraints," *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019.

[95] G. Zhai, L. Liu, L. Zhang, Y. Liu, and Y. Jiang, "Poseconvgru: A monocular approach for visual ego-motion estimation by learning," *Pattern Recognition*, vol. 102, 2020.

[96] J. Huang, S. Yang, T.-J. Mu, and S.-M. Hu, "Clustervo: Clustering moving instances and estimating visual odometry for self and surroundings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Conference Proceedings, pp. 2168–2177.

[97] B. Ummenhofer, H. Z. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," *30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017)*, pp. 5622–5631, 2017.

[98] Y. Wang and Y. F. Xu, "Unsupervised learning of accurate camera pose and depth from video sequences with kalman filter," *Ieee Access*, vol. 7, pp. 32 796–32 804, 2019.

[99] D. Barnes, W. Maddern, G. Pascoe, and I. Posner, "Driven to distraction: Self-supervised distractor learning for robust monocular visual odometry in urban environments," *2018 Ieee International Conference on Robotics and Automation (Icra)*, pp. 1894–1900, 2018.

[100] G. Iyer, J. K. Murthy, G. Gupta, K. M. Krishna, and L. Paull, "Geometric consistency for self-supervised end-to-end visual odometry," *Proceedings 2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition Workshops (Cvprw)*, pp. 380–388, 2018.

[101] Q. Liu, R. H. Li, H. S. Hu, and D. B. Gu, "Using unsupervised deep learning technique for monocular visual odometry," *Ieee Access*, vol. 7, pp. 18 076–18 088, 2019.

[102] T. Shen, Z. Luo, L. Zhou, H. Deng, R. Zhang, T. Fang, and L. Quan, *arXiv preprint arXiv:.09103*.

[103] N. Radwan, A. Valada, and W. Burgard, "Vlocnet++: Deep multitask learning for semantic visual localization and odometry," *Ieee Robotics and Automation Letters*, vol. 3, no. 4, pp. 4407–4414, 2018.

[104] S. Li, X. Wang, Y. Cao, F. Xue, Z. Yan, and H. Zha, "Self-supervised deep visual odometry with online adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Conference Proceedings, pp. 6339–6348.

[105] B. Ummenhofer, H. Z. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox, "Demon: Depth and motion network for learning monocular stereo," *30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017)*, pp. 5622–5631, 2017.

[106] S. Vijayanarasimhan, S. Ricco, C. Schmid, R. Sukthankar, and K. Fragkiadaki, "Sfm-net: Learning of structure and motion from video," *arXiv preprint arXiv:.07804*, 2017.

[107] J. Zhang, Q. Su, P. Liu, C. Xu, and Y. Chen, "Unsupervised learning of monocular depth and ego-motion with space–temporal-centroid loss," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 3, pp. 615–627, 2020.

[108] W. Zhao, S. Liu, Y. Shu, and Y.-J. Liu, "Towards better generalization: Joint depth-pose learning without posenet," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Conference Proceedings, pp. 9151–9161.

[109] W. Y. Ci, Y. P. Huang, and X. Hu, "Stereo visual odometry based on motion decoupling and special feature screening for navigation of autonomous vehicles," *Ieee Sensors Journal*, vol. 19, no. 18, pp. 8047–8056, 2019.

[110] J. Galarza, E. Pérez, E. Serrano, A. Tapia, and W. G. Aguilar, "Pose estimation based on monocular visual odometry and lane detection for intelligent vehicles," in *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*. Springer, Conference Proceedings, pp. 562–566.

[111] C. Zhao, L. Sun, Z. Yan, G. Neumann, T. Duckett, and R. Stolkin, "Learning kalman network: A deep monocular visual odometry for on-road driving," *Robotics and Autonomous Systems*, vol. 121, 2019.

[112] A. Jacobson, F. Zeng, D. Smith, N. Boswell, T. Peynot, and M. Milford, "Semi-supervised slam: Leveraging low-cost sensors on underground autonomous vehicles for position tracking," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Conference Proceedings, pp. 3970–3977.

[113] M. Bloesch, J. Czarnowski, R. Clark, S. Leutenegger, and A. J. Davison, "Codeslam-learning a compact, optimisable representation for dense visual slam," *2018 Ieee/Cvf Conference on Computer Vision and Pattern Recognition (Cvpr)*, pp. 2560–2568, 2018.

[114] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Conference Proceedings, pp. 12 716–12 725.

[115] A. Al-Kaff, D. Martin, F. Garcia, A. de la Escalera, and J. M. Armingol, "Survey of computer vision algorithms and applications for unmanned aerial vehicles," *Expert Systems with Applications*, vol. 92, pp. 447–463, 2018.

[116] J. Goppert, S. Yantek, and I. Hwang, "Invariant kalman filter application to optical flow based visual odometry for uavs," *2017 Ninth International Conference on Ubiquitous and Future Networks (Icufn 2017)*, pp. 99–104, 2017.

[117] A. Khan and M. Hebert, "Multi modal pose fusion for monocular flight with unmanned aerial vehicles," *2018 Ieee Aerospace Conference*, 2018.

[118] X. J. Yan, Z. Y. Shi, and Y. S. Zhong, "Vision-based global localization of unmanned aerial vehicles with street view images," *2018 37th Chinese Control Conference (Ccc)*, pp. 4672–4678, 2018.

[119] I. Cvisic, J. Cesic, I. Markovic, and I. Petrovic, "Soft-slam: Computationally efficient stereo visual simultaneous localization and mapping for autonomous unmanned aerial vehicles," *Journal of Field Robotics*, vol. 35, no. 4, pp. 578–595, 2018.

[120] Y. Liu and C. Q. Wang, "Hybrid real-time stereo visual odometry for unmanned aerial vehicles," *Optical Engineering*, vol. 57, no. 7, 2018.

[121] S. W. Yang, S. A. Scherer, X. D. Yi, and A. Zell, "Multi-camera visual slam for autonomous navigation of micro aerial vehicles," *Robotics and Autonomous Systems*, vol. 93, pp. 116–134, 2017.

[122] F. Bellavia, M. Fanfani, and C. Colombo, "Selective visual odometry for accurate auv localization," *Autonomous Robots*, vol. 41, pp. 133–143, 2017.

[123] M. Ferrera, J. Moras, P. Trouvé-Peloux, and V. Creuze, "Real-time monocular visual odometry for turbid and dynamic underwater environments," *Sensors (Switzerland)*, vol. 19, no. 3, 2019.

[124] M. M. Nawaf, D. Merad, J. P. Royer, J. M. Boï, M. Saccone, M. B. Ellefi, and P. Drap, "Fast visual odometry for a low-cost underwater embedded stereo system," *Sensors (Switzerland)*, vol. 18, 2018.

[125] Y. Cheng, M. W. Maimone, and L. Matthies, "Visual odometry on the mars exploration rovers - a tool to ensure accurate driving and science imaging," *Ieee Robotics and Automation Magazine*, vol. 13, no. 2, pp. 54–62, 2006.

[126] K. C. Di, F. L. Xu, J. Wang, S. Agarwal, E. Brodyagina, R. X. Li, and L. Matthies, "Photogrammetric processing of rover imagery of the 2003 mars exploration rover mission," *Isprs Journal of Photogrammetry and Remote Sensing*, vol. 63, no. 2, pp. 181–201, 2008.

[127] R. X. Li, S. W. Squyres, R. E. Arvidson, B. A. Archinal, J. Bell, Y. Cheng, L. Crumpler, D. J. D. Marais, K. Di, T. A. Ely, M. Golombek, E. Graat, J. Grant, J. Guinn, A. Johnson, R. Greeley, R. L. Kirk, M. Maimone, L. H. Matthies, M. Malin, T. Parker, M. Sims, L. A. Soderblom, S. Thompson, J. Wang, P. Whelley, and F. L. Xu, "Initial results of rover localization and topographic mapping for the 2003 mars exploration rover mission," *Photogrammetric Engineering and Remote Sensing*, vol. 71, no. 10, pp. 1129–1142, 2005.

[128] S. Zaman, L. Comba, A. Biglia, D. R. Aimonino, P. Barge, and P. Gay, "Cost-effective visual odometry system for vehicle motion control in agricultural environments," *Computers and Electronics in Agriculture*, vol. 162, pp. 82–94, 2019.

[129] D. Scaradozzi, S. Zingaretti, and A. J. S. C. Ferrari, "Simultaneous localization and mapping (slam) robotics techniques: a possible application in surgery," *Shanghai Chest*, vol. 2, no. 1, 2018.

[130] G. Dimas, D. K. Iakovidis, G. Ciuti, A. Karargyris, and A. Koulaouzidis, "Visual localization of wireless capsule endoscopes aided by artificial neural networks," *2017 Ieee 30th International Symposium on Computer-Based Medical Systems (Cbms)*, pp. 734–738, 2017.

[131] G. Dimas, D. K. Iakovidis, A. Karargyris, G. Ciuti, and A. Koulaouzidis, "An artificial neural network architecture for non-parametric visual odometry in wireless capsule endoscopy," *Measurement Science and Technology*, vol. 28, no. 9, 2017.

[132] D. K. Iakovidis, G. Dimas, G. Ciuti, F. Bianchi, A. Karargyris, A. Koulaouzidis, and E. Toth, "Robotic validation of visual odometry for wireless capsule endoscopy," *2016 Ieee International Conference on Imaging Systems and Techniques (Ist)*, pp. 83–87, 2016.

[133] E. Spyrou, D. K. Iakovidis, S. Niafas, and A. Koulaouzidis, "Comparative assessment of feature extraction methods for visual odometry in wireless capsule endoscopy," *Computers in Biology and Medicine*, vol. 65, pp. 297–307, 2015.

[134] F. A. Cheein, N. Lopez, C. M. Soria, F. A. di Sciascio, F. L. Pereira, and R. Carelli, "Slam algorithm applied to robotics assistance for navigation in unknown environments," *J Neuroeng Rehabil*, vol. 7, p. 10, 2010.

[135] P. Mountney, D. Stoyanov, A. Davison, and G. Z. Yang, "Simultaneous stereoscope localization and soft-tissue mapping for minimal invasive surgery," *Med Image Comput Comput Assist Interv*, vol. 9, no. Pt 1, pp. 347–54, 2006.

[136] P. Mountney and G. Z. Yang, "Motion compensated slam for image guided surgery," *Med Image Comput Comput Assist Interv*, vol. 13, no. Pt 2, pp. 496–504, 2010.

[137] J. Wei, G. Ye, T. Mullen, M. Grundmann, A. Ahmadyan, and T. Hou, "Instant motion tracking and its applications to augmented reality," 2019.

[138] R. Frikha, R. Ejbali, and M. Zaied, "Camera pose estimation for augmented reality in a small indoor dynamic scene," *Journal of Electronic Imaging*, vol. 26, no. 5, p. 053029, 2017.

[139] T. Gao and W. Jiang, "Monocular camera tracking curve optimization algorithm in augmented reality," in *International Conference on Broadband and Wireless Computing, Communication and Applications*. Springer, Conference Proceedings, pp. 295–303.

[140] W. Fang, L. Y. Zheng, and X. Y. Wu, "Multi-sensor based real-time 6-dof pose tracking for wearable augmented reality," *Computers in Industry*, vol. 92-93, pp. 91–103, 2017.

[141] A. Ivan, H. Seok, J. Lim, K.-J. Yoon, I. Cho, and I. K. Park, "Visual-inertial rgb-d slam for mobile augmented reality," in *Pacific Rim Conference on Multimedia*. Springer, Conference Proceedings, pp. 928–938.

[142] P. L. Li, T. Qin, B. T. Hu, F. Y. Zhu, and S. J. Shen, "Monocular visual-inertial state estimation for mobile augmented reality," *Proceedings of the 2017 Ieee International Symposium on Mixed and Augmented Reality (Ismar)*, pp. 11–21, 2017.

[143] X. Yang, J. B. Guo, T. L. Xue, and K. T. Cheng, "Robust and real-time pose tracking for augmented reality on mobile devices," *Multimedia Tools and Applications*, vol. 77, no. 6, pp. 6607–6628, 2018.

[144] E. J. Shamwell, K. Lindgren, S. Leung, and W. D. Nothwang, "Unsupervised deep visual-inertial odometry with online error correction for rgb-d imagery," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2478–2493, 2020.

[145] M. Abolfazli Esfahani, H. Wang, K. Wu, and S. Yuan, "Aboldeepio: A novel deep inertial odometry network for autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 5, pp. 1941–1950, 2020.

[146] L. Han, Y. Lin, G. Du, and S. Lian, "Deepvio: Self-supervised deep learning of monocular visual inertial odometry using 3d geometric constraints," pp. 6906–6913, 2019.

[147] C. Chen, S. Rosa, Y. Miao, C. X. Lu, W. Wu, A. Markham, and N. Trigoni, "Selective sensor fusion for neural visual-inertial odometry," pp. 10 534–10 543, 2019.

[148] F. Xue, X. Wang, S. Li, Q. Wang, J. Wang, and H. Zha, "Beyond tracking: Selecting memory and refining poses for deep visual odometry," pp. 8567–8575, 2019.

[149] X. Kuo, C. Liu, K. Lin, and C. Lee, "Dynamic attention-based visual odometry," pp. 160–169, 2020.